ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Relating compound toxicity to molecular structure using machine learning

Master Thesis

Robin Bosshard, 16-915-399

October 16, 2023

Supervisors: Prof. Dr. Fernando Perez-Cruz, Dr. Eliza Harris, Lili Gasser (SDSC)
Dr. Kasia Arturi (Eawag)

Department of Computer Science, ETH Zürich

Abstract

Abstract goes here.

Acknowledgments

First and foremost, I would like to thank Prof. Dr. Fernando Perez-Cruz from the Swiss Data Science Center (SDSC) for granting me the opportunity to work on this fascinating project. His support has been invaluable.

I would like to express my sincere gratitude to my supervisor Dr. Kasia Arturi from Swiss Federal Institute of Aquatic Science and Technology (Eawag) and my supervisors Dr. Eliza Harris, Lili Gasser from SDSC for their numerous discussions, patience and valuable insights. Without their help, this thesis would not have been achievable.

Additionally, I would like acknowledge Prof. Dr. Juliane Hollender from Eawag for her support throughout the project and for the enlightening experience of visiting the Eawag labs.

Furthermore, my gratitude goes out to Jason Brown, Feshuk Madison, and Katie Paul Friedman for their active participation in discussions concerning the technical aspects of the tcpl pipeline and the ToxCast database.

Lastly, I extend a special thank you to my family and friends for their unconditional support throughout my academic journey.

Contents

Chapter 1

Introduction

1.1 The Challenge of Environmental Pollution

Over the past few decades, the upsurge in environmental pollution by chemical compounds has been driven by industrial processes, agricultural methods, our consumerism and various other contributing factors. This has resulted in significant ecological and health issues. Although these chemicals are integral for many products and have the potential to improve our comfort of modern society, they can also pose risks and adversely affect both our health and the environment, either acutely or chronically. Toxic substances threaten wildlife but also makes our air, soil and finally our drinking water and food supply less safe. The EU currently maintains comprehensive chemical regulations, however, it is anticipated that global chemicals production will double by 2030 [1]. Moreover, the widespread utilization of chemicals, including their inclusion in consumer goods, is expected to expand further. Table A.2 provides an overview of omnipresent water pollutants. Even though there are over 275 million known chemical compounds registered by the Chemical Abstracts Service (CAS) [2], merely a tiny fraction of them undergo close monitoring via target analytical approaches and even less is known about their toxicity profiles and negative health effects on our organsims.

Building upon the European Green Deal [3], the 8th Environment Action Programme, guiding European environmental policy until 2030, reinforces the EU's goal of sustainable living within planetary limits, with a vision extending to 2050. One of its key objectives is a zero-pollution commitment, covering air, water, and soil, prioritizing the well-being of EU citizens. In particular, the European Commission published a sustainability-focused chemicals strategy (CSS), aligning with the EU's zero-pollution ambition with one of the objectives to minimize concerning substances by either substituting or phasing them out wherever feasible [4]. Consequently, the urgent need to monitor and effectively assess the hazards associated with

the daily entering of thousands of poorly understood chemicals into our environment becomes increasingly evident.

1.2 The Imperative for Prioritization and Toxicity Assessment

Modern analytical methods, especially high-resolution mass spectrometry (HRMS/MS), are becoming increasingly important in fields like metabolomics, drug discover, forensics and environmental science and toxicology. Nontarget HRMS/MS has improved the ability to detect emerging compounds in environmental samples, often with unknown toxicity profiles. These compounds are assessed based on factors such as abundance and fragmentation data. See in Figure 1.1 for an overview. However, the endeavor to identify compounds and characterize their toxicity remains a resource-intensive and time-consuming process. This challenge is further impeded by the scarcity of well-characterized substances that can be used as references for comparison when analyzing unknown compounds, hindering comprehensive elucidation. Traditionally, the prioritization of unidentified compounds rely on signal intensity as a guiding metric. Unfortunately, this approach falls short in delivering an accurate assessment of environmental exposures, as it tends to overlook the crucial toxicological dimension. Consequently, substances with the potential for severe ecological consequences, such as endocrine-disrupting compounds, frequently evade detection due to their low abundance, despite their high toxicity. Therefore, there is an urgent need for alternative hazard-driven prioritization strategies of unidentified NTS HRMS/MS signals that incorporate the toxicity and ecological impact more effectively.

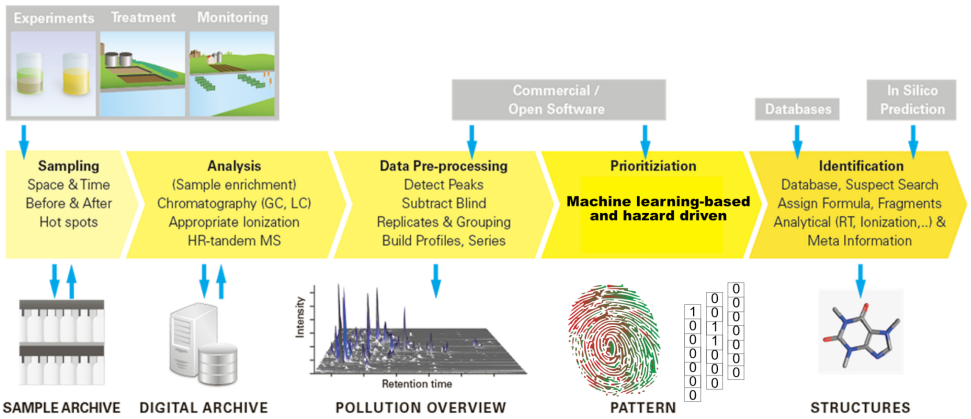


Figure 1.1: Figure 1 adapted (with modified Prioritization step) from Hollender et al. [5]: Nontarget screening with high resolution mass spectrometry in the environment: Ready to go?

1.3 The Promise of Machine Learning in Toxicity Prediction

In the past few years, machine learning has emerged as a transformative force in the field of toxicology, particularly in the realm of high-throughput toxicity prediction. High-throughput screening (HTS) has revolutionized the way we assess toxicity by allowing thousands of in vitro bioassays to be conducted rapidly. This high-throughput approach, coupled with advancements in robotics and automated analysis, has generated large volumes of toxicity data, paving the way for more comprehensive assessments of chemical compounds. Alongside the rise of machine learning, this advancement has facilitated the creation of predictive models capable of forecasting compound toxicity based on their chemical structure. These models can be trained on extensive datasets containing well-documented toxicity information, allowing them to learn the underlying patterns and relationships between chemical structure and target toxicity. Once trained, these models can reliably predict the toxicity of new compounds, even if they have not undergone laboratory testing. This approach holds the potential to significantly reduce the time and cost associated with early-stage toxicity pre-assessment and plays a crucial role in prioritizing compounds for further in-depth testing.

1.4 MLinvitroTox: A Novel Approach

In response to the pressing need for a more hazard-driven and comprehensive assessment of environmental contaminants, Arturi et al. introduced MLinvitroTox [6], an innovative machine learning framework. In particular it is the primary goal of this thesis to collaborate with the authors in further advancing and developing this framework. MLinvitroTox leverages molecular fingerprints extracted from fragmentation spectra [1], marking a fundamental shift in how we forecast the toxicity of the myriad unidentified HRMS/MS features. While traditional QSAR models predict bioactivities based on molecular fingerprints derived from chemical structures, MLinvitroTox was trained with supervised classification models on molcular fingerprints from chemical structures but is applied to molecular fingerprints generated from experimentally measured MS2 spectra using *SIRIUS* and *CSI:FingerID*. SIRIUS is a software package for annotating small molecules from nontarget HRMS/MS data, while CSI:FingerID is a machine-learning tool employed by SIRIUS to predict molecular fingerprints from fragmentation spectra. MLinvitroTox leverages streamlined machine learning techniques to predict the compounds bioactivity, respectively toxicity, ensuring a broad toxicological coverage encompassing nearly 300 target-specific and 90 cytotoxic endpoints, sourced from ToxCast/Tox21 data. Subsequently, the toxicity predictions generated

[1] also termed as Tandem mass spectrometry or MS/MS or MS2

by the framework are employed to prioritize compounds, with the flexibility to emphasize specific aspects of toxicity profiles tailored to individual preferences. This prioritization strategy facilitates more streamlined and thorough evaluations of environmental contaminants, enhancing a more hazard-driven risk assessment.

1.5 Objectives and Significance

The central objective of this thesis is to develop a streamlined framework for the prediction of compound toxicity across multiple endpoints, resulting in the creation of toxicity fingerprint. The generated toxicity fingerprints will provide valuable insights for the prioritization process in identifying most hazardous compounds found in environmental samples, ultimately contributing to the preservation of ecosystems and our health. The framework aims to develop a custom curation of structural and toxicological data to address challenges from modeling heterogenuous, and imbalanced data sets. Notably, the use of SIRIUS molecular fingerprints and xgboost (Extreme Gradient Boosting) models, complemented by feature selection?, has yielded consistently successful results. Furthermore, we have validated the effectiveness of MLinvitroTox by applying it to MassBank spectra, demonstrating an average balanced accuracy of 0.75? in predicting toxicity.

1.6 Thesis Structure

The initial chapters lay the groundwork by providing essential background information and summarizing related work. As we progress through the subsequent chapters, we will delve into the methodology and technical intricacies involved in preparing ToxCast/Tox21 toxicity data, transforming them into suitable inputs for our machine learning pipeline. This foundational work will serve as the cornerstone for the forthcoming chapters, where we will showcase the potential of MLinvitroTox. Additionally, will also demonstrate the framework's effectiveness through validation using real-world data and discuss about the implications of our research.

Chapter 2

Background

This chapter provides background information necessary to understand the rest of the thesis. We introduce the ToxCast's invitro database together with processing pipeline tcpl to get familiar how the bioactivity labels are generated from the in vitro high-throughput screening (HTS) data. We also introduce the concept of molecular fingerprints that are used as the features for the machine learning models.

2.1 The Evolution of Toxicity Testing

Back in 2007, the U.S. National Academy of Sciences introduced a visionary perspective and published a landmark report, titled as "Toxicity Testing in the 21st Century: Vision and Strategy", advocating a transition from conventional, time-consuming animal-based in vivo tests to efficient high-throughput in vitro pathway assays on cells.

In the realm of high-throughput screening (HTS), a multitude of in vitro bioassays that improve chemical screening can be executed, thanks to the advancements in robotics, data processing and automated analysis. This synergy yields to the generation of extensive toxicity datasets like ToxCast and Tox21.

HTS datasets like ToxCast and other sources opened the door to promising applications of machine learning in predictive computational toxicology. We can develop predictive models to screen environmental chemicals that have little toxicity data available, where the outcomes can be used for further testing prioritization. Such models often forcast toxicity based on chemical structures using Quantitative Structure-Activity Relationships (QSARs) and molecular fingerprints. Molecular fingerprints encode molecules as fixed-length binary vectors, denoting the presence (1) or absence (0) of specific substructures or functional groups.

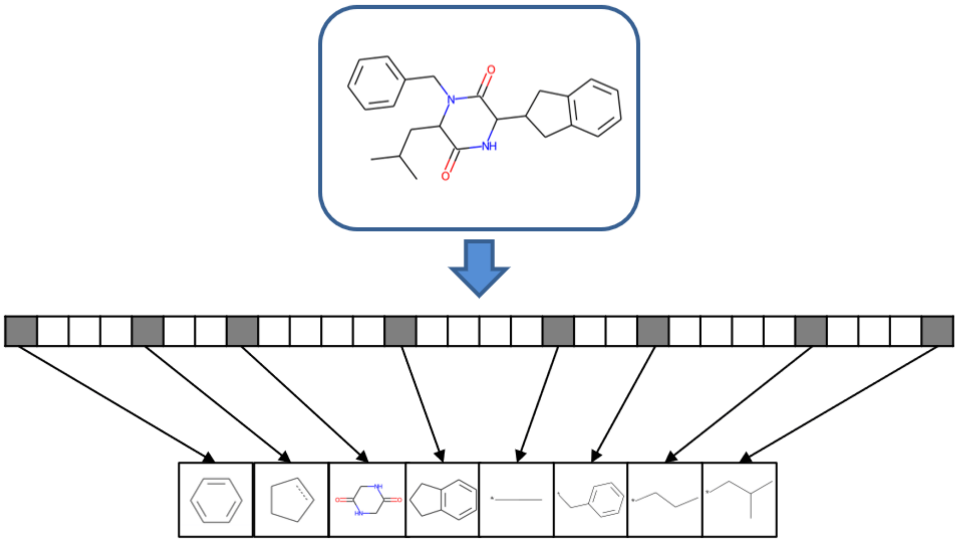


Figure 2.1: Figure 1 adapted from Janela et al (2022) [7]. Schematic molecular fingerprint. Each bit position accounts for the presence or absence of a specific structural fragment. Bit positions are set on (set to 1, gray) if the substructure is present in a molecule, or set off (white) if it is absent.

The utilization of molecular fingerprints for in vitro toxicity prediction is based on the assumption that molecular toxic effects result from relatively straightforward interactions between distinct chemical components and receptors during a molecular initiating event (MEI). On a larger biological scale, the MEI can set in motion a sequential chain of causally linked key events (KE) at different levels of biological organisation from within cells to potentially culminating in an adverse outcome pathway (AOP) at the organ or organism level, as depicted in Figure X. The mechanistic information captured in AOPs reveal how chemicals or stressors cause harm, offering insights into disrupted biological processes, potential intervention points but also guide regulatory decisions on next generation risk assessment and precise toxicity testing. With the AOP framework we have a analytical construct that allows an activity mapping from the presence or absence of certain molecular substructures encoded in the fingerprints to target mechanistic toxicity. Finally, when monitoring disruptions in toxicity pathways, pharmacokinetic models can be leveraged to extrapolate in vitro findings to human blood and tissue concentrations.

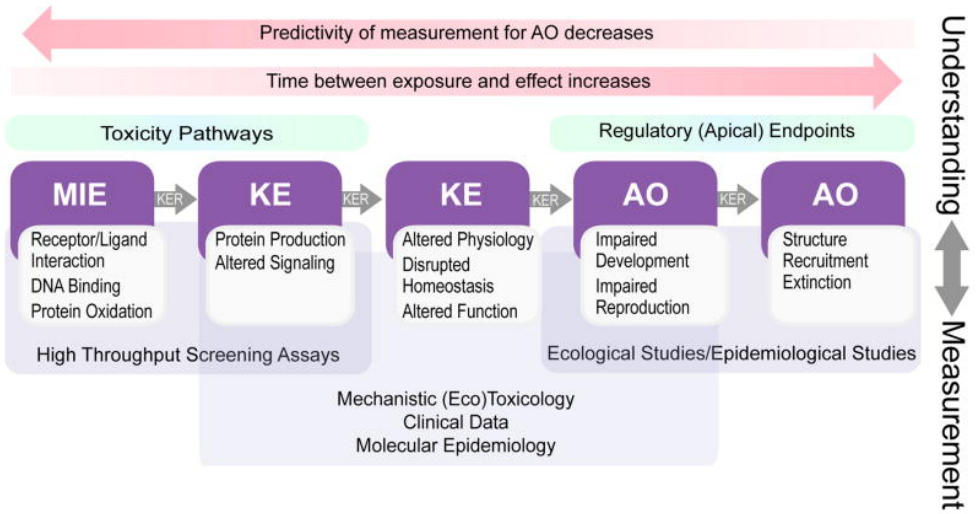


Figure 2.2: Figure 1 from From Ankley and Edwards (2018): Harvesting the promise of AOPs: An assessment and recommendations [8]. Depiction of the role of the adverse outcome pathway (AOP) framework. MIE=Molecular Initiating Event, KE=Key Event, KER=Key Event Relationship, AO=Adverse Outcome

2.2 InvitroDB v4.1

The most recent release of the ToxCast's (Toxicity Forecaster) database [1], referred to as invitroDBv4.1, serves as a source of an extensive collection of high-throughput screening (HTS) targeted bioactivity data. This database encompasses information on a total of 10 196 compounds, selectively screened across 1485 assay endpoints. The assays utilize a range of technologies to assess the impact of chemical exposure on a wide array of biological targets, including individual proteins and cellular processes such as mitochondrial health, developmental processes and nuclear receptor signaling.

This resource originated from the collaboration of two prominent institutions: the United States Environmental Protection Agency[2] (EPA) through its ToxCast program and the National Institutes of Health[3] (NIH) via the Tox21[4] initiative. Utilizing data gathered from various research laboratories, this relational database is publicly available and can be downloaded[5] by visiting the official ToxCast website.

[1] released on September 21, 2023
[2] https://www.epa.gov
[3] https://ntp.niehs.nih.gov/whatwestudy/tox21
[4] https://tox21.gov/
[5] https://www.epa.gov/chemical-research/exploring-toxcast-data

2.3 tcpl v3.0

In chapter 1, we introduce the Python re-implementation *pytcpl* of the core components of the ToxCast pipeline *tcpl*, originally an R package. The tcpl package offers a comprehensive suite of tools for managing HTS data, provides consistent and reproducible concentration-response modeling and populates the MYSQL database, invitrodb. The multiple-concentration screening paradigm intends to pinpoint the activity of compounds, while also estimating their efficacy and potency. The concentration-response modeling procedure also addresses outlier robustness and signal loss due to cytotoxicity.

To streamline cross-experiment comparisons and reduce parameter complexity, concentration-response modeling adheres to a zero-centered, positive response paradigm. Negative response data undergoes inverse transformation during normalization. To ensure robustness without data exclusion, a log-likelihood function utilizing a Student's t-distribution with 4 degrees of freedom [9] is employed. The model with the lowest Akaike Information Criterion (AIC) value is selected as the *winning* model. The winning model is then used to estimate the efficacy and potency of the compound. The potency estimates, also called point-of-departure (POD) estimates, are derived from the fitted curve characteristics, identifying concentrations at which the model curve crosses certain response levels. For example, the activity concentration at which the compound reaches 50% of its maximum response is denoted by *ac50* and similarly the activity concentration at cutoff efficacy by *acc*. Addionionally, the package calculates assay noise by computing the median absolute deviation over response values from the first two concentrations (bmad). The baseline region is defined as $0 + 3bmad$, and *acb* is the concentration where the model first reaches 3bmad. The figure illustrates the four POD estimates. To classify[6] a concentration-response series as an active hit and allowing to determine PODs, the following criteria must be met: The *winning* model must be either the Hill or gain-Loss model, with modeled curve's peak surpassing the assay-specific efficacy cutoff, and at least one concentration should have a median response exceeding this threshold. Notably, no POD estimates are computed when the compound is considered inactive[7], as these estimates are not applicable in such cases.

2.3.1 Tcplfit2

To improve upon tcpl, R package *tcplfit2* was developed, a standalone package focused on curve-fitting and hit-calling. The package also offers a more flexible and robust fitting procedure, allowing for the use of different opti-

[6] legacy tcpl employs only binary classification
[7] if the *winning* fit model was the constant model

mization algorithms and the incorporation of user-defined constraints. (Todo: Explain MLE, Compare Strictly standardized mean difference.) The package also includes a more comprehensive set of POD estimates, including the *ac10* and *ac95* estimates. Tcplfit2 differs from other R-language open-source concentration-response packages like *drc* and *mixtox*, as it is specifically tailored for HTS concentration-response data, offering an extensive set of curve models, summarized in Table **??**.

Table 2.1: tcplfit2 Model Details

Model	**Label**	**Equations**[1]
Constant	cnst	$f(x) = 0$
Linear	poly1	$f(x) = ax$
Quadratic	poly2	$f(x) = a\left(\frac{x}{b} + \left(\frac{x}{b}\right)^2\right)$
Power	pow	$f(x) = ax^p$
Hill	hill	$f(x) = \frac{tp}{1+\left(\frac{ga}{x}\right)^p}$
Gain-Loss	gnls	$f(x) = \frac{tp}{\left(1+\left(\frac{ga}{x}\right)^p\right)\left(1+\left(\frac{x}{la}\right)^q\right)}$
Exponential 2	exp2	$f(x) = a\left(\exp\left(\frac{x}{b}\right) - 1\right)$
Exponential 3	exp3	$f(x) = a\left(\exp\left(\left(\frac{x}{b}\right)^p\right) - 1\right)$
Exponential 4	exp4	$f(x) = tp\left(1 - 2^{-\frac{x}{ga}}\right)$
Exponential 5	exp5	$f(x) = tp\left(1 - 2^{-\left(\frac{x}{ga}\right)^p}\right)$

[1] Model parameters: a: x-scale, b: y-scale p: (gain) power, q: (loss) power, tp: top, ga: gain AC50, la: loss AC50

All the models to assume that the normalized observations from the concentration-response series are not normally distributed but follow a Student's t-distribution with four degrees of freedom. The Student's t-distribution has heavier tails compared to the normal distribution, making it more robust to outlier and eliminates the necessity of removing potential outliers prior to the fitting process. The model fitting algorithm in `tcplFit2` employs maximum likelihood estimation to estimate model parameters for all available models.

Consider $t(z, \nu)$ as the Student's t-distribution with ν degrees of freedom, where y_i represents the observed response for the i-th observation, and μ_i is the estimated response for the same observation. The calculation of z_i is as follows:

$$z_i = y_i - \mu_i \exp(\sigma)$$

where σ is the scale term.

Then the log-likelihood is

$$\sum_{i=1}^{n} [\ln(t(z_i, 4)) - \sigma]$$

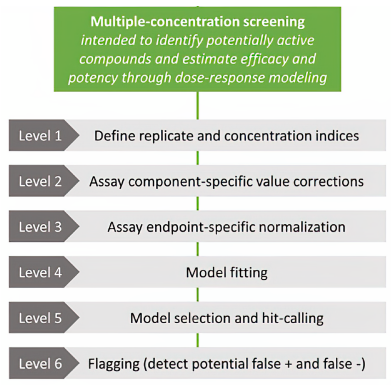
where n is the number of observations.

Hitcalling:

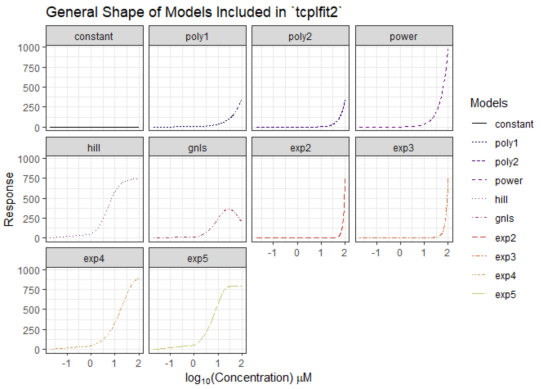
The `tcplhit2_core` continuous hitcall is calculated based on the product of the probabilities of the following values: (i) that at least one median response is greater than the cutoff; (ii) that the top of the fitted curve is above the cutoff, and (iii) that the winning AIC value is less than that of the constant model. The first probability is computed by using the error parameter from the model fit and t-distribution to calculate the odds of at least one response exceeding the cutoff (the error model around the data uses a 3-parameter t-distribution). The second is by using the likelihood ratio to compute the one-sided probability of the cutoff being exceeded. The third is set to be the Akaike weight relative to the constant model:

$$\frac{e^{-\frac{1}{2}AIC_{\text{winning}}}}{e^{-\frac{1}{2}AIC_{\text{winning}}} + e^{-\frac{1}{2}AIC_{\text{cnst}}}}.$$

Multiple-concentration processing includes six processing levels. Briefly, level 1 processing defines concentration and replicate indices, giving integer values 1…N to increasing concentrations and technical replicates, where 1 represents the lowest concentration or first technical replicate. Level 2 processing allows for basic transformations of the raw data, e.g. logarithmic conversion, and removes data deemed poor quality by the user. Similar to level 1 in single-concentration processing, level 3 normalizes data to fold-change or percent-of-control and converts concentrations to a logarithm scale. At level 4 data are modeled (described below), before level 5 processing defines the winning model and the activity call. Level 5 processing also separates each data series into categories to facilitate easy triaging of the results. Level 6 processing identifies potential false positive and false negative results, giving problematic data series a flag.



(a) Overview of tcpl's data analysis pipeline for multiple concentrations screening data.



(b) Employed curve-fit models in tcpl v3.0 for fitting concentration-response data series through the application of maximum likelihood estimation.

Figure 2.3: tcpl

2.4 Chemical Target Toxicity vs. Cytotoxicity

Intuitively, we expect increasing chemical concentrations to result in increasing bioactivity. However, this is not always the case. At higher doses the chemical can become cytotoxic leading to dying cells. In consequently, a reduction in bioactivity can occur, e.g., the activation of the reporter gene decreases.

Chemical toxicity can manifest in diverse ways, falling into two major categories [10]:

1. **Specific toxicity** is the result of a chemical's interaction and disruption of a specific biomolecular target or pathway, such as a receptor agonist/antagonist effect or enzyme activation/inhibition.
2. the **Cytotoxicity and cell stress** is the generalized disruption of the cellular machinery. Cell-disruptive processes encompass various mechanisms, such as protein, DNA, or lipid reactivity, or processes like apoptosis, oxidative stress responses or mitochondrial disruption. Cell viability can be evaluated either individually or concurrently. One approach is to assess it by calculating the proportion of live cells in a population, employing a fluorescent dye that specifically enters living cells. This dye remains incapable of permeating the membranes of deceased cells, resulting in fluorescence intensity directly correlating with cell viability.

It is common for compounds to exhibit target bioactivity within a limited concentration range, which coincides with a nonspecific activation response in the presence of cell stress and cytotoxicity. Figure 2.4 illustrates the interference between specific toxicity and cytotoxicity.

A related phenomenon is referred to as the *cytotoxicity burst* [10], where e.g., the reporter genes can be non-specifically induced close to cell death [11]. The Toxcast pipeline is designed to minimize false negatives due to the inclusive risk assessment.Nevertheless, attributed to interference processes, the reliability of reported activities becomes uncertain and false positives are possible without a cytotoxicity evaluation. While a portion of the assay activity within this concentration range might indeed reflect chemical interactions with the intended assay target, another portion does not.

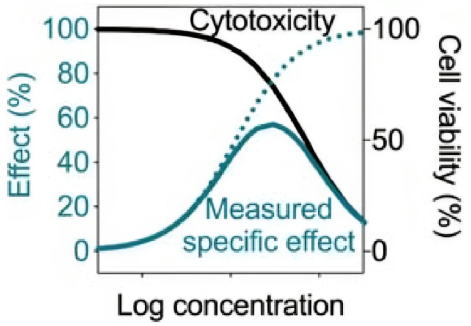


Figure 2.4: Figure 7.8 from Escher at al. [11]: Bioanalytical Tools in Water Quality Assessment: Second Edition. Example of a bioassay response with cytotoxicity interference. The dotted line shows the theoretical effect but due to cytotoxicity (black line is cell viability), the measured effect has an inverted U-shape. The measured effect can addionally be confounded and intensified by the cytotoxicity burst, where even an exponential shape is likely for the gaining part. In this case, the effect should be only evaluated up to some concentration.

2.5 Challenges of HTS data

2.6 Molecular Fingerprints

Chapter 3

Related work

A recent review highlights the proliferation of research employing invitroDB since 2006, encompassing topics such as assessing chemical toxicity, identifying contaminants for environmental monitoring, and computational toxicity forecasting. The majority of ML applications based on invitroDB have predominantly concentrated on specific target endpoints and cytotoxicity. Notably, research has extensively covered adverse outcomes related to endocrine receptor systems, including androgen and estrogen receptors, alongside areas such as carcinogenicity, hepatic steatosis, hepatotoxicity, immunotoxicity, developmental toxicity, neurotoxicity, and cardiotoxicity.

Typically, various mathematical models or curve shapes are employed to analyze the data for the best fit. Several commercial tools and open-source libraries are available for this purpose. One widely used system for managing high-throughput screening (HTS) concentration-response data is tcpl (ToxCast Pipeline), but also.

Compared to similar efforts in the field where ecotoxicity was predicted from MS2 based on in vivo data, in the current work, the invitroDB toxicity database was used to train supervised classification models for hundreds of available toxicity endpoints

In their systematic investigation using Tox21 data, Wu et al. (2021) explored the impact of various modeling approaches and chemical features on predictive toxicology, with a focus on model performance and explainability trade-offs. The study found that the assay endpoint from the Tox21 data being predicted was the most significant factor influencing model performance. Endpoints with higher predictability, characterized by lower data imbalance and larger datasets, performed well regardless of the modeling approach or molecular representation. For less predictable endpoints, simpler models like Linear Regression performed similarly to complex ones, prioritizing both predictivity and interpretability. Moreover this study suggests consensus

modeling and multi-task learning to enhance predictability and model performance across endpoints. In this thesis, we set the goal to not overlook simpler models due to their higher interpretability and comparable performance. As suggested we do not further investigate on the differnt molecular representations and use a fixed compilation of molecolar fingerprints[1] as initial input features. We incorporated in our studies a form of consensus modeling to consolidate predictability and multi-task learning to improve model performance across different endpoints.

[1]SIRIUS

Chapter 4

Material and Methods

4.1 Data Overview

Presence Matrix

Consider a collection of m assay endpoints, denoted by $A = \{a_1, a_2, \ldots, a_m\}$ and a set of n compounds represented as $C = \{c_1, c_2, \ldots, c_n\}$. To facilitate data comprehension, we introduce a *presence matrix* $P \in \{0,1\}^{m \times n}$. Rows, indexed by i, represent assay endpoints a_i, while columns, indexed by j, denote presence (1) or absence (0) of compound c_j in those endpoints. Matrix P is sparse due to the selective testing of compounds across different assay endpoints. A compound is considered present in an assay endpoint if it has undergone testing and a corresponding concentration-response series is available. See Figure 4.1 for a visual of the *presence matrix P* covering all assay endpoints and compounds in *invitroDBv3.5*.

Subsetting data

We exclusively consider assay endpoints that have been tested with a minimum of 2000 compounds. This criterion ensures the availability of sufficient data for the training of a machine learning model. Refer to Figure 4.2 for a visual representation of the *presence matrix P*, which now encompasses only the resulting subset of all assay endpoints within *invitroDBv3.5*. From now on, we will call this specific subset the data that we will be focusing on for this thesis.

Concentration-Response Series

A *concentration-response series* is represented as a set of k concentration-response pairs:

$$S = \{(conc_1, resp_1), (conc_2, resp_2), \ldots, (conc_k, resp_k)\}$$

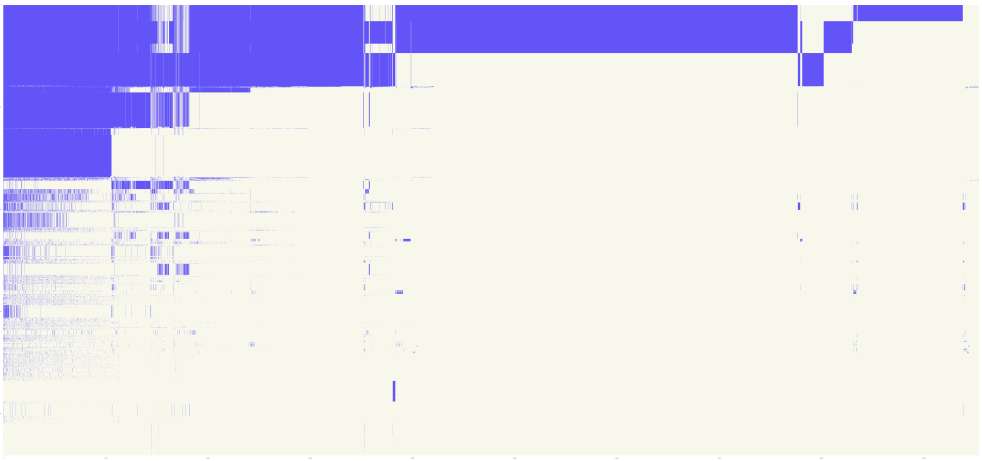


Figure 4.1: The *presence matrix* P covering all assay endpoints and compounds available in *invitroDBv3.5* with $m = 2205$ assay endpoints and $n = 9541$ compounds. The presence matrix is organized by sorting it based on the number of compounds present in each assay endpoint and the compounds are arranged in descending order of their presence frequency. The total count, where $P_{ij} = 1$, indicates the availability of 3 342 377 concentration-response series for downstream analysis.

For each entry in the presence matrix P with $P_{ij} = 1$, we collect the corresponding concentration-response series S_{ij} for the compound c_j in the assay endpoint a_i. We analyse in total $\sum_{i,j} P_{ij} = 1\,372\,225$ concentration-response series, comprising a sum of $\sum_{i,j} |S_{ij}| = 48\,861\,036$ concentration-response pairs across all compounds and assay endpoints. We get the concentration-response pairs by combining tables mc0, mc1, and mc3 from invitroDBv3.5. We also gather necessary sample information such as well type, row, and column index from the assay well-plate. The concentrations are transformed to the logarithmic scale using the unit μM (micromolar), while the responses are normalized to either fold-induction or percent-of-control units. Figure 4.3 showcases a single concentration-response series for some compound tested within an assay endpoint.

In this section, we demonstrate the significance of variations in concentration-response pairs among different compounds and assay endpoints. In practice, concentrations are often subjected to multiple testing iterations, resulting in the formation of distinct concentration groups. Within each concentration group, the number of replicates is indicated by n_{rep}. We introduce the following quantities corresponding to a concentration-response series for a compound c_i in a given assay endpoint a_i:

- $n_{\text{datapoints}_{i,j}}$: the total number of concentration-response pairs ($|S|$)
- $n_{\text{groups}_{i,j}}$: the number of distinct concentrations tested

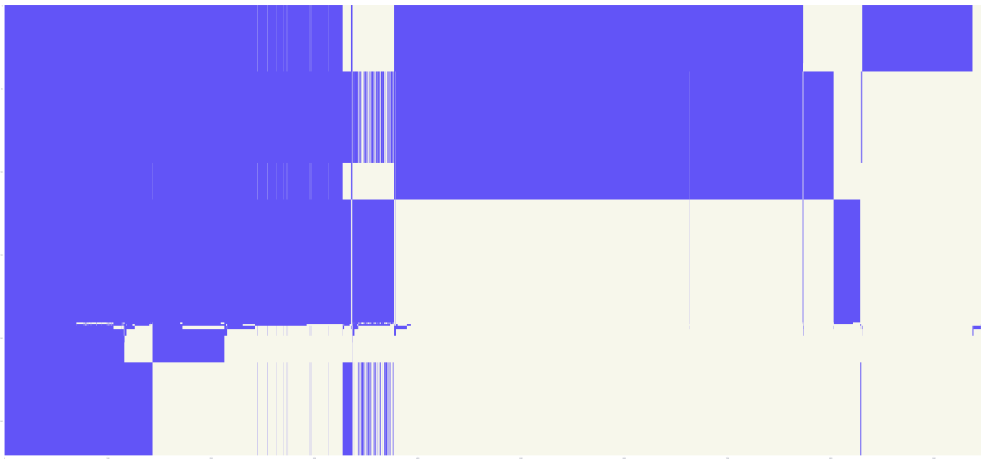


Figure 4.2: The *presence matrix* P covering only the subset of all of assay endpoints available in *invitroDBv3.5*, considered for this thesis, encompassing $m = 271$ assay endpoints and $n = 9456$ compounds. The total count, where $P_{ij} = 1$, indicates the availability of 1 372 225 concentration-response series for downstream analysis.

Figure 4.3: A concentration-response series for the compound *Estropipate* (DTXSID3023005) in the assay endpoint *TOX21_ERa_LUC_VM7_Agonist* (aeid=788). The series has a total of $k = 45$ concentration-response pairs and is composed of $n_{conc} = 15$ concentration groups, each with $n_{rep} = 3$ replicates.

- $n_{\text{replicates}_{i,j}}$: the number of replicates for each concentration group
- $min_{\text{conc}_{i,j}}$: the lowest concentration tested
- $max_{\text{conc}_{i,j}}$: the highest concentration tested

For an overview of these quantities across the entire set of considered concentration-response series, please refer to Figure 4.4. This figure illustrates the above metrics aggregated by their means, grouped by assay endpoints

and compounds.

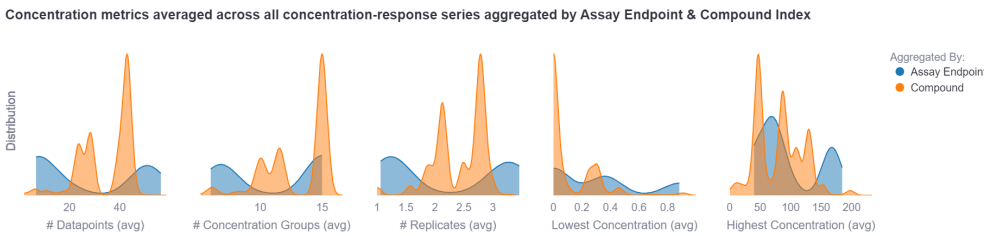


Figure 4.4: Concentration metrics averaged across all concentration-response series aggregated by assay endpoint (blue) and compound (orange). E.g., the first chart shows the distribution on the average number of datapoints across all assay enpoint $a_i \in A$ with $\frac{1}{|C|}\sum_j n_{\text{datapoints}_{i,j}}$ and across all compounds $c_j \in C$ with $\frac{1}{|A|}\sum_i n_{\text{datapoints}_{i,j}}$. The same is done for the other metrics: $n_{\text{groups}_{i,j}}$, $n_{\text{replicates}_{i,j}}$, $min_{\text{conc}_{i,j}}$, and $max_{\text{conc}_{i,j}}$.

4.2 Pytcpl

We introduce pytcpl, a streamlined Python package inspired by the R package tcpl, designed for processing high-throughput screening data. The package primarily focuses on providing essential features such as concentration-response curve fitting and allows for continuous hit-calling for compound bioactivity across diverse assay endpoints, akin to tcplfit2. Invitrodb version 3.5 release can optionally serve as backend database if desired. The package optimizes data storage and provides compressed raw data and metadata from *invitroDB* in Parquet files. This efficient strategy reduces storage needs, resulting in just 4 GB within the repository—compared to the original 80 GB database. This obviates the need for a cumbersome, large-scale database installation, rendering downstream analysis more accessible and efficient. Our package is crafted to accomodate cusomizable processing steps and facilitate interactive data visualization with curve surfer. Moreover, it empowers Python-oriented researchers to seamlessly engage in data analysis and exploration.

4.2.1 Pipeline

1. Data collection
2. Cutoff determination and filtering (Meet conditions for curve fitting)
3. Curve fitting
4. Hit calling

Data Collection

First, all datapoints are collected from the database and assigned to the concentration response-series belonging to the respective compound in the corresponding assay endpoint.

Curve Fitting

Introduce all candidate fit models, discuss the pros and cons of each model. Discuss the fitting procedure, how the models are fitted, Maximum Likelihood Estimation.

Table 4.1: tcplfit2 Model Details

Model	Label	Equations[1]	Role in pytcpl
Exponential 3	exp3	$f(x) = a\left(\exp\left(\left(\frac{x}{b}\right)^p\right) - 1\right)$	Omit
Gain-Loss 2	gnls2	$f(x) = \frac{tp}{1+\left(\frac{ga}{x}\right)^p} \exp\left(-qx\right)$	New

[1] Model parameters: a: x-scale, b: y-scale p: (gain) power, q: (loss) power, tp: top, ga: gain AC50

Hit Calling

Akaike criterion, probability of being active, etc..

$$AIC = -2\log(L(\hat{\theta}, y)) + 2K \tag{4.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.2}$$

4.2.2 Curve Surfer

Data visualization, overview of what is possible with the tool. Filter by assay endpoint, compound, etc.

4.3 Machine Learning Pipeline

4.3.1 Preprocessing

Subselecting the columns from the output tables generated by pytcpl: DTXSID identifier and continuous hitcall value. The feature inputs to the machine learning model is a molecular structure represented as fingerprint generated from a SMILES string uniquely determined by the compounds DTXSID

identifier. The SMILES string is a linear representation of a compound's molecular structure. The SMILES string is converted to a molecular graph, which is then converted to a feature vector. The feature vector is then used to train a machine learning model. The machine learning model is then used to predict the hitcall value for a given compound. The machine learning pipeline is illustrated in Figure **??**.

4.3.2 Binary Classification

The goal is to predict whether a compound is active or inactive for a given assay endpoint. We can formulate this as a binary classification problem, where the input is the compound's molecular structure fingerprint and the output is the hitcall value binarized by some decision threshold. The hitcall value is rendered to a binary variable, where 1 indicates that the compound is active and 0 indicates that the compound is inactive.

4.3.3 Regression

4.3.4 Massbank Validation

Chapter 5

Results and Discussion

5.1 Results

5.2 Evaluation

5.3 Discussion

Chapter 6

Conclusion

We have evidence of a multitude of chemicals being present in the environment and in our bodies and that mixture exposure indeed matters. This knowledge needs to be deepened, and the quantitative contribution of chemicals to compromised health should be better described and translated into regulatory action. As indicated in a scientific opinion paper of the German Federal Environmental Agency (Conrad et al. 2021), the CSS goals may be considered as a moving target. For increasing scientific evidence and improved method for detection and assessment of chemicals, development of new technologies require innovative regulatory, technological and societal reactions. We should be flexible and prepared to take up the scientific challenges and collaborate productively with regulatory institutions to address the identified challenges and modernise chemical risk assessment. This is also in line with the concern of many scientists that chemical pollution and the wide range of adverse effects on human and ecosystem health demand additional efforts on a global scale (Brack et al. 2022; Wang et al. 2021). We see the CSS as a European strategy that, in concert with other initiatives, may open new opportunities to minimise hazardous chemical pollution and thus risks to human health and ecosystems.

Bibliography

[1] U. N. E. Programme, *Global chemicals outlook ii - from legacies to innovative solutions: Implementing the 2030 agenda for sustainable development - synthesis report*, 2019. [Online]. Available: https://wedocs.unep.org/20.500.11822/27651.

[2] C. A. Service, *Chemical abstracts service (cas) is a division of the american chemical society*, Source of chemical information located in Columbus, Ohio, United States, https://www.cas.org/support/documentation/cas-databases, 2023.

[3] E. Commission, D.-G. for Research, and Innovation, *European Green Deal - Research & innovation call*. Publications Office of the European Union, 2021. DOI: 10.2777/33415.

[4] E. Commission, "Eu chemicals strategy for sustainability towards a toxic-free environment," 2020, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Chemicals Strategy for Sustainability Towards a Toxic-Free Environment. [Online]. Available: https://environment.ec.europa.eu/strategy/chemicals-strategy_en.

[5] J. Hollender, E. L. Schymanski, H. P. Singer, and P. L. Ferguson, "Nontarget screening with high resolution mass spectrometry in the environment: Ready to go?" *Environmental Science & Technology*, vol. 51, no. 20, pp. 11 505–11 512, 2017, PMID: 28877430. DOI: 10.1021/acs.est.7b02184. eprint: https://doi.org/10.1021/acs.est.7b02184. [Online]. Available: https://doi.org/10.1021/acs.est.7b02184.

[6] K. Arturi and J. Hollender, "Machine learning-based hazard-driven prioritization of features in nontarget screening of environmental high-resolution mass spectrometry data," *Environmental Science & Technology*, vol. 0, no. 0, null, 0, PMID: 37279189. DOI: 10.1021/acs.est.3c00304.

eprint: https://doi.org/10.1021/acs.est.3c00304. [Online]. Available: https://doi.org/10.1021/acs.est.3c00304.

[7] T. Janela, K. Takeuchi, and J. Bajorath, "Introducing a chemically intuitive core-substituent fingerprint designed to explore structural requirements for effective similarity searching and machine learning," *Molecules*, vol. 27, no. 7, 2022, ISSN: 1420-3049. DOI: 10.3390/molecules27072331. [Online]. Available: https://www.mdpi.com/1420-3049/27/7/2331.

[8] G. T. Ankley and S. W. Edwards, "The adverse outcome pathway: A multifaceted framework supporting 21st century toxicology," *Current Opinion in Toxicology*, vol. 9, pp. 1–7, 2018, Risk assessment in Toxicology, ISSN: 2468-2020. DOI: https://doi.org/10.1016/j.cotox.2018.03.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468202017301420.

[9] "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2290063 (visited on 09/20/2023).

[10] R. Judson *et al.*, "Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space," *Toxicological Sciences*, vol. 152, no. 2, pp. 323–339, May 2016, ISSN: 1096-6080. DOI: 10.1093/toxsci/kfw092. eprint: https://academic.oup.com/toxsci/article-pdf/152/2/323/26290632/kfw092.pdf. [Online]. Available: https://doi.org/10.1093/toxsci/kfw092.

[11] B. Escher, P. Neale, and F. Leusch, *Bioanalytical Tools in Water Quality Assessment*. IWA Publishing, Jun. 2021, ISBN: 9781789061987. DOI: 10.2166/9781789061987. eprint: https://iwaponline.com/book-pdf/899726/wio9781789061987.pdf. [Online]. Available: https://doi.org/10.2166/9781789061987.

[12] R. Schwarzenbach *et al.*, "The challenge of micropollutants in aquatic systems," *Science (New York, N.Y.)*, vol. 313, pp. 1072–7, Sep. 2006. DOI: 10.1126/science.1127291.

Appendix A

Appendix

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Table A.1: Assay Source Names and Long Names

assay_source_name	assay_source_long_name
ACEA	ACEA Biosciences
APR	Apredica
ATG	Attagene
BSK	Bioseek
NVS	Novascreen
OT	Odyssey Thera
TOX21	Tox21/NCGC
CEETOX	Ceetox/OpAns
LTEA	LifeTech/Expression Analysis
VALA	VALA Sciences
CLD	CellzDirect
CCTE_PADILLA	CCTE Padilla Lab
TANGUAY	Tanguay Lab
STM	Stemina Biomarker Discovery
ARUNA	ArunA Biomedical
CCTE	CCTE Labs
CCTE_SHAFER	CCTE Shafer Lab
CPHEA_STOKER	CPHEA Stoker and Laws Labs
CCTE_GLTED	CCTE Great Lakes Toxicology and Ecology Division
UPITT	University of Pittsburgh Johnston Lab
UKN	University of Konstanz
ERF	Eurofins
TAMU	Texas A&M University
IUF	Leibniz Research Institute for Environmental Medicine
CCTE_MUNDY	CCTE Mundy Lab
UTOR	University of Toronto, Peng Laboratory

Table A.2: Table 2 adapted from [12]. Examples of ubiquitous water pollutants.

Origin/Usage	Class	Selected Examples	Related Problems
Industrial Chemicals	Solvents	Tetrachloromethane	Drinking-water contamination
	Intermediates	Methyl-t-butylether	
	Petrochemicals	BTEX (benzene, toluene, xylene)	
Industrial Products	Additives	Phthalates	
	Lubricants	PCBs (polychlorinated biphenyls)	Biomagnification, long-range transport
	Flame Retardants	Polybrominated diphenylethers	
Consumer Products	Detergents	Nonylphenol ethoxylates	Endocrine active transformation product (nonylphenol)
	Pharmaceuticals	Antibiotics	Bacterial resistance, nontarget effects
	Hormones	Ethinyl estradiol	Feminization of fish
Biocides	Pesticides	DDT	Toxic effects and persistent metabolites
	Non-agricultural biocides	Tributyltin	Endocrine effects
Geogenic/ Natural Chemicals	Heavy Metals	Lead, cadmium, mercury	
	Inorganics	Arsenic, selenium, fluoride, uranium	Risks for human health, Drinking-water-quality
	Taste and Odor	Geosmin, methylisoborneol	
	Human Hormones	Estradiol	Feminization of fish
Disinfection/ Oxidation	Disinfection by-products	Trihalomethanes, haloacetic acids, bromate	Drinking-water–quality, human health problems
Transformation Products	Metabolites from all above	Metabolites of perfluorinated compounds	Bioaccumulation despite low hydrophobicity
		Chloroacetanilide herbicide metabolites	Drinking-water–quality